Outline of Geographic Data Structure Issues in TRANSIMS

Brian Bush and Rob Oakes

Los Alamos National Laboratory

4 August 1994

PURPOSE

The purpose of this presentation is to make the TRANSIMS team members aware of the major issues which may influence the choice of geographic data structures in TRANSIMS and to gather feedback from the team members concerning their requirements for these data structures.

INTRODUCTION

- By *geographic data structure* we mean a way organizing, storing, and accessing geographic data.

- Exactly what the geographic data in TRANSIMS is remains to be specified.

- Two important aspects of a data structure used for storing persistent data are its *internal* (memory) and its *external* (disk) representations. These two representations need not be identical, but they must be consistent with each other. It is also conceivable that more than one internal or external representation may be implemented.

  - An *internal* representation takes the form of a set of C++ classes with member functions for accessing and manipulating the data.
  - An *external* representation takes the form of a database where the data is stored permanently and from which it can be retrieved.

GOALS

We have identified several goals for the geographic data structures:

- They must *meet* the geographic data *requirements* of the planner, microsimulation, and analyst's toolbox.
- They must *be efficient* in terms of access speed, storage required, and querying capabilities.
- They must *be portable*—able to work on multiple hardware platforms and in all of the TRANSIMS software.
- They must *be general* enough to avoid any duplication of effort in writing software and to eliminate any need for supplementary databases or ad-hoc additions to the database.

MAJOR ISSUES

- The most important issues related to the choice of the geographic data structures are:
  - What geographic data is required for TRANSIMS?
  - What are the most efficient data structures for TRANSIMS?
  - In what architectural configuration will the data structures reside?
- Secondary issues—those which must be considered carefully, but which are not the primary driving force behind decision making—are:
  - What sort of data structures and GIS/CAD packages are currently in use by the transportation community?
  - What standards are there for geographic data and which organizations are developing standards?
  - What commercial software is available and would it be desirable to use it for TRANSIMS? Also, is there work at LANL we can build on?

GEOGRAPHIC DATA

- There are several types of data in TRANSIMS which may (or may not) be considered “geographic” and be included in the geographic data structures. What the data are depends on the detailed requirements of the various components of TRANSIMS.
  - basic geographic data for road segments (path with latitude, longitude, and elevation)
  - supplementary road segment attributes (grade, curvature, number of lanes, class, lane width, speed limit, etc.)
  - basic geographic data for junctions/intersections (location and connected road segments)
  - supplementary junction/intersection attributes (characteristics, signaling, visibility, turning patterns, etc.)
  - vehicle position and state as a function of time
  - traveler location as a function of time
  - elevation model
  - location of administrative areas (census blocks, census tracks, city boundaries, county boundaries, etc.)
  - demographic/economic data tied to census blocks or tracts
  - urban area specification (land use, building heights, etc.)
  - origin-destination data
  - trip plans

GEOGRAPHIC DATA (continued)

- The question here is what partition should be made between geographic and non-geographic data, if any, and how the two should be separated. (Most GISs, for example, have the ability to handle both types of data.)

- Thus we need to determine what the fundamental *objects* in TRANSIMS are and which of these objects are to be considered geographic objects. This determination must be based on the data requirements of the planner, microsimulation, and analyst's toolbox.

EFFICIENCY OF GEOGRAPHIC DATA STRUCTURES

- The efficiency and the applicability of a spatial data structure depends on the uses that will be made of it. Structures that are efficient for display are not necessarily efficient for queries, and those that are efficient for intersection queries are not efficient for containment queries.

- The three major classes of spatial data structures (grids, *k*-trees, and *k*d-trees) each have advantages. In order to choose the best data structures we need to know to requirements for data access, visualization, and spatial and temporal querying. This will depend on the functionality required for the planner, microsimulation, analyst's toolbox, and applications.

- We cannot separate the question of the internal representation from the external one because the quantity of geographic data required to describe a large city is greater than can fit into memory. Thus efficient disk access and organization (paging) must be studied.

GEOGRAPHIC QUERIES

- Several types of query operators may have to be supported in TRANSIMS, particularly in the analyst's toolbox, but also in the planner and microsimulation.
  - *General operators:* intersection, union, complement, difference.
  - *Geographic operators:* point search, range search, near, nearest, adjacent to, containment, intersection.
  - *Temporal operators:* before, after, between, near.
- Here are examples of possible TRANSIMS queries:
  - "Find out which road segment (or vehicle) is nearest to where the user pressed the mouse."
  - "Display the difference between the flow volumes in all commercial areas between the morning and afternoon rush hours"
  - "Display all of the destinations of the cars owned by people living in census blocks with average income over a certain amount."
  - "Load all of the road and intersections in a given part of the city into the microsimulator."

GEOGRAPHIC DATA STRUCTURE TAXONOMY

- There are four major types of geographic data structures.

Representations	*Description*	*Efficiency*
Sweep	data ordered along axis	good for nearest neighbors bad for intersection searches
Grid	data sorted by location in grid	good for display bad for range searches
k-Tree	extends binary tree in multiple dimensions	good for certain search large storage requirements
*k*d-Tree	binary tree with alternating dimension index	good for certain search bad for parallel implementation

- Implementations vary in the way that they deal with the following:
  - indexing (radix vs. variable)
  - treatment of objects with spatial extent
  - single vs. multiple references to an object
  - disk paging schemes
  - insertion and deletion methods
  - balancing

ARCHITECTURAL CONSIDERATIONS

- The database can be either centralized on a server, distributed over a network, or reside in multiple local copies made from a central archive. It may have to be accessible from a variety of platforms.

- Different TRANSIMS components will deal with data differently, so the data structures should be flexible and general.

  - In order for the microsimulation to execute rapidly, the data will have to reside in memory rather than on disk. The exact form of the representation in memory will probably depend upon whether a simulation is running on a massively parallel, parallel, networked, or single-cpu machine. Also, the strictly geographic data (i.e., coordinates as opposed to attributes such as number of lanes) is not necessary for the microsimulation
  - The applications, the planner, and the analyst's toolbox will probably only need indices in memory to the data on disk, with appropriate caching.

GIS/CAD USAGE IN THE TRANSPORTATION COMMUNITY

- Metropolitan Planning Organizations and other parts of the transportation community use several GIS/CAD packages.
  - Some use GIS/CAD packages specialized for transportation systems analysis (GIS/Trans, GIS-T, and TransCAD).
  - Others use transportation system analysis software with GIS/CAD capabilities (McTrans, TRANPLAN, Transyt, etc.).
  - Still others use general GIS/CAD packages such as AutoCAD and ArcInfo.
- In general these software packages use their own data formats and integrate geographic and non-geographic information into a single database. Some of the formats will probably have to be supported by TRANSIMS.

GEOGRAPHIC DATA STRUCTURE STANDARDS

- It is important to know what the existing standards in the field are for two reasons:
  - It will probably be necessary to support some of the data formats currently in use by transportation organizations.
  - We may want to adopt one of the standards ourselves or at least use one as an organizational guide.

- Several standards organizations are developing standards for spatial data. Most of the committees have not published draft specifications yet and will not have final specifications for several years.
  - Within ANSI, the Geographic Database Committee (X3L1) had it first meeting in January 1994; members include ESRI ArcInfo, Intergraph, GeoVision, Unisys, System 9, Geographic Data Technology, and ACE. The ANSI committee for multimedia SQL (SQL/MM) is considering geographic extensions to SQL and the ANSI committee developing the object-oriented SQL3 intends to include spatial objects in its standard.
  - ISO has a Geographic Information/Geomatics Standards group.
  - The Object Database Management Group (ODMG, part of OMG, the Object Management Group) has been asked by X/Open to coordinate its work on data structures with ANSI X3.

GEOGRAPHIC DATA STRUCTURE STANDARDS (continued)

- The Federal Geographic Data Committee (FGDC) has a Standards Working Group. Its has started a National Geospatial Data Clearinghouse based on its Spatial Metadata Standard. The FGDC is part of the National Spatial Data Infrastructure (NSDI) which is part of the National Information Infrastructure (NII), administrated by NIST.
- OpenGIS has published a draft of the Open Geodata Interoperability Specification (OGIS). Members include the U.S. Army Corps of Engineers (USACOE), NASA, Intergraph, PCI Remote Sensing, and the Center for Advanced Spatial Technology (CAST). The standard includes a Virtual Geodata Model (VGM) and an Applications Programming Model (APM) and is oriented toward support of SQL3 with multimedia spatial extensions.
- There are two older U.S. Government standards called the Spatial Data Transfer Standard (SDTS) and the Spatial Archive and Interchange Format (SAIF).
- The DoD has MIL-STDs for both raster and vector graphics, but the file formats are rather cumbersome.

- There are also a number of *de facto* standards currently in use:
  - AutoCAD DXF file format is used by many organizations, including transportation organizations, for CAD drawings.
  - ArcInfo and MapInfo file formats are widely used in the GIS community and also by several transportation organizations.
  - A number of the U.S. Geological Survey (USGS) formats (such as DLG and DEM) have become *de facto* standards.
  - The U.S. Bureau of the Census TIGER/Line file format is also widely used in the transportation community.

POSSIBLE ROLES FOR COMMERCIAL SOFTWARE

- If existing software can meet our requirements and constraints, it may be advantageous to purchase it because of the advantages in terms of development time, testing, functionality, and user support.

- Important constraints related to commercial software are cost, licensing, continued support, platform dependence, integrability, interoperability, and sponsor approval.

- There are three classes of software to be considered:

  – *Database.* Types of databases of possible interest are relational databases (RDBs), object databases (ODBs), and tree-based databases (TDBs). Some products have built-in support for geographic and temporal data.
  – *GIS/CAD.* These products handle display, storage, and analysis of the geographic data. Although they often use custom data structures, they generally can be integrated with external databases. Major vendors are Intergraph, ESRI ArcInfo, MapInfo, and AutoDesk.
  – *Class Libraries.* Commercially available class libraries are available to support database and graphics programming, but probably do not provide specific support for spatial data structures.

Conclusion

- Because of the large number of issues connected to choosing geographic data structures, it is necessary to prioritize our investigation and research. Below is a list of what we presently see as the top priorities:

  - identification of the geographic objects and their interaction
  - determination of how to partition the geographic and non-geographic data, if at all
  - relative efficiency of spatial data structures for access, visualization, spatial queries, temporal queries
  - unification of internal with external structures and among internal structures in different platforms and programs
  - hardware, cost, transportation community, and time constraints
  - requirements for planner, microsimulation, toolbox, applications